SOLVING THEMATIC DIVERGENCES IN MACHINE TRANSLATION

Bonnie Dorr*
M.I.T. Artificial Intelligence Laboratory
545 Technology Square, Room 810
Cambridge, MA 02139, USA
internet: bonnie@reagan.ai.mit.edu

ABSTRACT

Though most translation systems have some mechanism for translating certain types of divergent predicate-argument structures, they do not provide a general procedure that takes advantage of the relationship between lexical-semantic structure and syntactic structure. A divergent predicate-argument structure is one in which the predicate (*e.g.*, the main verb) or its arguments (*e.g.*, the subject and object) do not have the same syntactic ordering properties for both the source and target language. To account for such ordering differences, a machine translator must consider language-specific syntactic idiosyncrasies that distinguish a target language from a source language, while making use of lexical-semantic uniformities that tie the two languages together. This paper describes the mechanisms used by the UNITRAN machine translation system for mapping an underlying lexical-conceptual structure to a syntactic structure (and *vice versa*), and it shows how these mechanisms coupled with a set of general linking routines solve the problem of thematic divergence in machine translation.

1 INTRODUCTION

There are a number of different divergence types that arise during the translation of a source language to a target language. Figure 1 shows some of these divergences with respect to Spanish, English, and German.[1]

We will look at each of these traditionally difficult divergence types in turn. The first divergence type is a structural divergence in that the verbal object is realized as a noun phrase (*John*) in English and as a prepositional phrase (*a Juan*) in Spanish. The second divergence is conflational. Conflation is the incorporation of necessary participants (or arguments) of a given action. Here, English uses the single word *like* for the two German words *haben* (*have*) and *gern* (*likingly*); this is because the manner argument (*i.e.*, the *likingly* portion of the lexical token) is incorporated into the main verb in English. The third divergence type is a lexical divergence as illustrated in the *stab* example by the choice of a different lexical word *dar* (literally *give*) for the word *stab*. The fourth divergence type is categorial in that the predicate is adjectival (*hungry*) in English but nominal (*hunger*) in German. Finally, the fifth divergence type is a thematic divergence: the object (*Mary*) of the English sentence is translated as the subject (*María*) in the Spanish sentence.

Divergence Type	*Translation Example*
Structural	I saw John ⇕ Vi a Juan (I saw to John)
Conflational	I like Mary ⇕ Ich habe Marie gern (I have Mary likingly)
Lexical	I stabbed John ⇕ Yo le di puñaladas a Juan (I gave knife-wounds to John)
Categorial	I am hungry ⇕ Ich habe Hunger (I have hunger)
Thematic	I like Mary ⇕ María me gusta a mí (Mary pleases me)

Figure 1: Divergence Types in Machine Translation

The final divergence type, thematic divergence, is the one that will be the focus of this paper. We will look at

*This paper describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for this research has been provided by NSF Grant DCR-85552543 under a Presidential Young Investigator's Award to Professor Robert C. Berwick. Useful guidance and commentary during this research were provided by Bob Berwick, Noam Chomsky, Bruce Dawson, Ken Hale, Mike Kashket, Jeff Siskind, and Patrick Winston. The author is also indebted to three anonymous reviewers for their aid in reshaping this paper into its current form.

[1] Many sentences may fit into these divergence classes, not just the ones listed here. Also, a single sentence may exhibit any or all of these divergences.

how the UNITRAN system [Dorr, 1987, 1990] solves the thematic divergence problem by mapping an underlying lexical-conceptual structure to a syntactic structure (and *vice versa*) on the basis of a set of general linking routines and their associated mechanisms. The other divergences are also handled by the UNITRAN system, but these are discussed in [Dorr, 1990].

It turns out there are two types of thematic divergences that show up in the translation of a source language to a target language: the first type consists of a reordering of arguments for a given predicate; and the second type consists of a reordering of predicates with respect to their arguments or modifiers. We will look at examples of each of these types in turn.

In the first case, an example is the reversal of the subject with an object as in the English-Spanish example of gustar-like shown in figure 1. The predicate-argument structures are shown here:[2]

(1)
[I-MAX [N-MAX María]
[V-MAX [V-1 [V-MIN me gusta] [P-MAX a mí]]]]
⇕
[I-MAX [N-MAX I]
[V-MAX [V-1 [V-MIN like] [N-MAX Mary]]]]

Here the subject *María* has reversed places with the object *mí*. The result is that the object *mí* turns into the subject *I*, and the subject *María* turns into the object *Mary*. The reverse would be true if translation went in the opposite direction.

An example of the second case of thematic divergence (not shown in figure 1) is the *promotion* of a complement up to the main verb, and the *demotion* of the main verb into an adjunct position (or *vice versa*). By *promotion*, we mean placement "higher up" in the syntactic structure, and by *demotion*, we mean placement "lower down" in the syntactic structure. This situation arises in the translation of the Spanish sentence *Juan suele ir a casa* into the English sentence *John usually goes home*:

(2)
[I-MAX [N-MAX Juan]
[V-MAX [V-1 [V-MIN suele]
[V-MAX [V-MIN ir] [P-MAX a casa]]]]]
⇕
[I-MAX [N-MAX John]
[V-MAX [V-1 [V-1 usually [V-MIN goes]]
[N-MAX home]]]]

Here the main verb *soler* takes *ir* as a complement; but, in English, the *ir* predicate has been placed into a higher position as the main verb *go*, and *soler* is placed into a lower position as the adjunct *usually* associated with the main verb. The reverse would be true if translation went in the opposite direction.

[2]Often times a native speaker of Spanish will invert the subject to post-verbal position:

[I-MAX e_i [V-MAX [V-1 [V-MIN me gusta] [P-MAX a mí]]]
[N-MAX María]$_i$].

However, this does not affect the internal/external reversal scheme described here since inversion takes place independently after thematic divergences have been handled.

Another example of the second case of thematic divergence is the demotion of the main verb into a complement position, and the promotion of an adjunct up to the main verb (or *vice versa*). This situation arises in the translation of the German sentence *Ich esse gern* into the English sentence *I like eating*:

(3)
[I-MAX [N-MAX Ich]
[V-MAX [V-1 [V-1 [V-MIN esse] gern]]]]
⇕
[I-MAX [N-MAX I]
[V-MAX [V-1 [V-MIN like] [V-MAX eating]]]]

Here the main verb *essen* takes *gern* as an adjunct; but, in English, *gern* has been placed into a higher position as the main verb *like*, and the *essen* predicate has been placed into a lower position as the complement *eating* of the main verb. The reverse would be true if translation went in the opposite direction.[3]

This paper will show how the system uses three mechanisms along with a set of general linking routines (to be defined) to solve thematic divergences such as those that have been presented. The next section introduces the terminology and mechanisms that are used in the solution of these divergences, and, in so doing, it will provide a brief glimpse of how thematic divergences are tackled. Section 3 discusses other approaches (and their shortcomings) in light of the thematic divergence problem. Finally, section 4 presents a general solution for the problem of thematic divergences, showing in more detail how a set of general linking routines and their associated mechanisms provide the appropriate mapping from source to target language.

2 TERMINOLOGY AND MECHANISMS

Before we examine thematic divergences and how they are solved, we must first look at the terminology and mechanisms used throughout this paper:[4]

[3]It might be argued that a "direct" translation is possible for each of these three examples:

(1′) Mary pleases me
(2′) John is accustomed to going home
(3′) I eat willingly

The problem with taking a direct approach is that it is not general enough to handle a wide range of cases. For example, *gern* can be used in conjunction with *haben* to mean *like*: *Ich habe Marie gern* ('I like Mary'). The literal translation, *I have Mary likingly*, is not only stylistically unattractive, but it is not a valid translation for this sentence. In addition, the direct-mapping approach is not bidirectional in the general case. Thus, even if we did take (1′), (2′), and (3′) to be the translations for (1), (2), and (3), we would not be able to apply the same direct mapping on the English sentences of (1), (2), and (3) (translating in the opposite direction) because we would still need to translate *like* and *usually* into Spanish and German. It is clear that we need some type of uniform method for translating thematic divergences.

[4]The terms complement, specifier, and adjunct have not been defined; roughly, these correspond to syntactic object,

Definition 1: An *LCS* is a lexical conceptual structure conforming to a modified version of Jackendoff's well-formedness rules [Jackendoff, 1983]. For example, *I like Mary* is represented as:

[$_{State}$ BE$_{Ident}$
([$_{Thing}$ REFERENT],
[$_{Place}$ AT$_{Ident}$
([$_{Thing}$ REFERENT], [$_{Thing}$ PERSON])],
[$_{Manner}$ LIKINGLY])]

Definition 2: An *RLCS* is an uninstantiated LCS that is associated with a root word definition in the lexicon (*i.e.*, an LCS with unfilled variable positions). For example, an RLCS associated with the word *like* is:

[$_{State}$ BE$_{Ident}$
([$_{Thing}$ X],
[$_{Place}$ AT$_{Ident}$ ([$_{Thing}$ X], [$_{Thing}$ Y])],
[$_{Manner}$ LIKINGLY])]

Definition 3: A *CLCS* is a composed (instantiated) LCS that is the result of combining two or more RLCS's by means of unification (roughly). This is the *interlingua* or language-independent form that is the pivot between the source and target language. For example if we compose the RLCS for *like* with the RLCS's for *I* ([$_{Thing}$ REFERENT]) and *Mary* ([$_{Thing}$ PERSON]), we get the CLCS corresponding to *I like Mary* (as shown in definition 1).

Definition 4: An *Internal Argument Position* is a syntactic complement for a lexical word of category V, N, A, P, I, or C.[5]

Definition 5: An *External Argument Position* is a syntactic specifier of N for a lexical word of category N or a specifier of I for a lexical word of category V.

Definition 6: An *Adjunct Argument Position* is a syntactic modifier that is neither internal nor external with respect to a lexical word.

Each word entry in the lexicon is associated with an RLCS, whose variable positions may have certain restrictions on them such as internal/external and promotion/demotion information (to be described). The CLCS is the structure that results from combining the lexical items of a source-language sentence into a single underlying pivot form.

The mapping that solves thematic divergences is defined in terms of the RLCS, the CLCS, the syntactic structure, and the markers that specify internal/external and promotion/demotion information. These markers, or *mechanisms*, are specified as follows:

Mechanism 1: The :INT and :EXT markers are override position markers that determine where the internal and external arguments will be positioned for a given lexical root word.

For example, the lexical entry for *gustar* is an RLCS that looks like the RLCS for *like* (see definition 2) except that it includes the :INT and :EXT markers:

[$_{State}$ BE$_{Ident}$
([$_{Thing}$ X :INT],
[$_{Place}$ AT$_{Ident}$ ([$_{Thing}$ X], [$_{Thing}$ Y :EXT])],
[$_{Manner}$ LIKINGLY])]

During the mapping from the CLCS (shown in definition 1) to the syntactic structure, the RLCS for *gustar* (or *like*) is matched against the CLCS, and the arguments are positioned according to the specification associated with the RLCS.[6] Thus, the :INT and :EXT markers account for the syntactic distinction between Spanish and English by realizing the [$_{Thing}$ REFERENT] node of the CLCS (corresponding to X in the RLCS) as the internal argument *mí* in Spanish, but as the external argument *I* in English; and also by realizing the [$_{Thing}$ PERSON] node of the CLCS (corresponding to Y in the RLCS) as the external argument *María* in Spanish, but as the internal argument *Mary* in English. Note that the :INT and :EXT markers show up only in the RLCS. The CLCS does not include any such markers as it is intended to be a language-independent representation for the source- and target-language sentence.

Mechanism 2: The :PROMOTE marker associated with an RLCS $\mathcal{H}$ places a restriction on the complement $\mathcal{P}1$ of the head $\mathcal{H}1$.[7] This restriction forces $\mathcal{P}1$ to be promoted in the CLCS as the head $\mathcal{P}$. $\mathcal{H}$ is then dropped into a modifier position of the CLCS, and the logical subject of $\mathcal{P}$ is inherited from the CLCS associated with the syntactic subject of $\mathcal{H}1$.[8]

For example, the lexical entry for *soler* contains a :PROMOTE marker that is associated with the RLCS: [$_{Manner}$ HABITUALLY :PROMOTE]

Thus, in the above formula $\mathcal{H}1$ corresponds to *soler*, and $\mathcal{P}1$ corresponds to the complement of *soler*. The :PROMOTE marker forces the syntactic complement $\mathcal{P}1$ to be promoted into head

subject, and modifier, respectively. For a more detailed description of these and some of the other definitions here, see [Dorr, 1990].

[5]V, N, A, P, I, and C stand for Verb, Noun, Adjective, Preposition, Inflection, and Complementizer, respectively.

[6]The lexical-selection procedure that maps the CLCS to the appropriate RLCS (for *like* or *gustar*) is not described in detail here (see [Dorr, 1990]). Roughly, lexical selection is a unification-like process that matches the CLCS to the RLCS templates in the lexicon, and chooses the associated lexical words accordingly.

position as $\mathcal{P}$ in the CLCS, and the head $\mathcal{H}$/ to be demoted into modifier position as $\mathcal{H}$ in the CLCS. So, in example (2) of the last section, the resulting CLCS is:[9]

[Event GO$_{Loc}$
  ([Thing PERSON],
   [Path TO$_{Loc}$
    ([Place AT$_{Loc}$ ([Thing PERSON], [Place HOME])])],
   [Manner HABITUALLY])]

Here the RLCS for *soler*, [Manner HABITUALLY], corresponds to $\mathcal{H}$ and the RLCS for *ir*, [Event GO ...], corresponds to $\mathcal{P}$. In the translation to English, [Manner HABITUALLY] is not promoted, so it is realized as an adjunct *usually* of the main verb *go*.

Mechanism 3: The :DEMOTE marker associated with an RLCS $\mathcal{P}$ places a restriction on the head $\mathcal{H}$/ of the adjunct $\mathcal{P}$/. This restriction forces $\mathcal{H}$ to be demoted into an argument position of the CLCS, and the logical subject of $\mathcal{P}$ to be inherited from the logical subject of $\mathcal{H}$.

For example, the lexical entry for *gern* contains a :DEMOTE marker that is associated with the Y argument in the RLCS:

[State BE$_{Circ}$
  ([Thing X],
   [Place AT$_{Circ}$ ([Thing X], [Event Y :DEMOTE])],
   [Manner LIKINGLY])]

Thus, in the above formula, $\mathcal{P}$/ corresponds to *gern* and $\mathcal{H}$/ corresponds to the syntactic head that takes *gern* as an adjunct. The :DEMOTE marker forces the head $\mathcal{H}$/ to be demoted into an argument position as $\mathcal{H}$ in the CLCS, and the adjunct $\mathcal{P}$/ to be promoted into head position as $\mathcal{P}$ in the CLCS. So in example (3) of the last section, the resulting CLCS is:

[State BE$_{Circ}$
  ([Thing REFERENT],
   [Place AT$_{Circ}$
    ([Thing REFERENT],
     [Event EAT ([Thing REFERENT], [Thing FOOD])])],
   [Manner LIKINGLY])][10]

Here the RLCS for *gern*, [State BE$_{Circ}$ ...], corresponds to $\mathcal{P}$ and the RLCS for *essen*, [State EAT ...], corresponds to $\mathcal{H}$. In the translation to English, [State BE$_{Circ}$ ...] is not demoted, so it is realized as the main verb *like* that takes *eating* as its complement.

[7] In general, a syntactic argument *u*/ is the canonical syntactic realization ($\mathcal{CSR}$) of the corresponding CLCS argument *u*. The $\mathcal{CSR}$ function is a modified version of a routine proposed in [Chomsky, 1986]. See [Dorr, 1990] for a more detailed discussion of this function.

[8] The logical subject is the highest/left-most argument in the CLCS.

Now that we have looked briefly at the mechanisms involved in solving thematic divergences in UNITRAN, we will look at how other approaches have attempted to solve this problem.

3 PREVIOUS APPROACHES

In tackling the more global problem of machine translation, many people have addressed different pieces of the thematic divergence problem, but no single approach has yet attempted to solve the entire space of thematic divergence possibilities. Furthermore, the pieces that *have* been solved are accounted for by mechanisms that are not general enough to carry over to other pieces of the problem, nor do they take advantage of cross-linguistic uniformities that can tie seemingly different languages together.

Gretchen Brown has provided a model of German-English translation that uses *lexical semantic structures* [Brown, 1974]. The work is related to the model developed for UNITRAN since both use a form of conceptual structure as the basis of translation. While this approach goes a long way toward solving a number of translation problems (especially compound noun disambiguation), it falls short of providing a systematic solution to the thematic divergence problem. This is largely because the conceptual structure does not serve as a common representation for the source and target languages. Instead, it is used as a point of transfer, and as such, it is forced to encode certain language-specific idiosyncrasies such as the syntactic positioning of conceptual arguments. In terms of the representations used in UNITRAN, this approach is analogous to using a language-to-language mapping from the RLCS's of the source language to the RLCS's of the target language without using an intermediate language-independent structure as a pivot form. In

[9] It should be noted that promotion and demotion structures are inverses of each other. Thus, although this CLCS looks somewhat "English-like," it is possible to represent the CLCS as something that looks somewhat "Spanish-like:"

[State Be$_{Circ}$
  ([Thing PERSON],
   [Place AT$_{Circ}$
    ([Thing PERSON],
     [Event GO$_{Loc}$
      ([Thing PERSON],
       [Path TO$_{Loc}$
        ([Place AT$_{Loc}$ ([Thing PERSON], [Place HOME])])])])],
   [Manner HABITUALLY])]

In this case, we would need to use the :DEMOTE marker (see mechanism 3) instead of the :PROMOTE marker, but this marker would be used in the RLCS associated with *usually* instead of the RLCS associated with *soler*. The justification for using the "English-like" version for this example is that the [Manner HABITUALLY] constituent is generally thought of as an aspectual element associated with a predicate (*e.g.*, in German, the sentence would be *Ich gehe gewöhnlich nach Hause* ('I go usually home')); this constituent cannot be used as a predicate in its own right. Thus, the complicated "Spanish-like" predicate-argument structure is not a likely conceptual representation for constructions that use [Manner HABITUALLY].

[10] The default object being eaten is [Thing FOOD], although this is not syntactically realized in this example.

this approach, there is no single language-independent mechanism that links the conceptual representation to the syntactic structure; thus, it is necessary to hand-code the rules of thematic divergence for English and German, and all divergence generalizations are lost.

In 1982, Lytinen and Schank developed the MOPTRANS Spanish-English system based on *conceptual dependency* networks [Lytinen & Schank, 1982].[11] This approach is related to the UNITRAN model of translation in that it uses an interlingual representation as the pivot from source to target language. The key distinction is that the approach lacks a generalized linking to syntax. For example, there is no systematic method for determining which conceptual argument is the subject and which is the object. This means that there is no uniform mechanism for handling divergences such as the subject-object reversal of example (1).

The LMT system is a logic-based English-German machine translator based on a modular logical grammar [McCord, 1989]. McCord specifically addresses the problem of thematic divergence in translating the sentence *Mir gefällt der Wagen* (*I like the car*). However, the solution that he offers is to provide a "transfer entry" that interchanges the subject and object positions. There are two problems with this approach. First it relies specifically on this object-initial ordering, even though the sentence is arguably more preferable with a subject-initial ordering *Der Wagen gefällt mir*; thus, the solution is dependent on syntactic ordering considerations, and will not work in the general case. Second the approach does not attempt to tie this particular type of thematic divergence to the rest of the space of thematic divergence possibilities; thus, it cannot uniformly translate a conceptually similar sentence *Ich fahre das Wagen gern* (*I like to drive the car*).

4 THEMATIC DIVERGENCES

In section 1, we introduced some examples of thematic divergences, and in section 2 we described some of the mechanisms that are used to solve these divergences. Now that we have looked at other machine translation approaches with respect to the thematic divergence problem, we will look at the solution that is used in the UNITRAN system.

Recall that there are two types of thematic divergences:

1. Different argument positionings with respect to a given predicate.
2. Different predicate positionings with respect to arguments or modifiers.

The first type covers the case of argument positions that diverge; it is accounted for by the :INT and :EXT markers. The second type covers the case of predicate positions that diverge; it is accounted for by the :PROMOTE and :DEMOTE markers. Together, these two types of divergences account for the entire space of thematic divergences, since all participants must be one of these two (either an argument, or a predicate, or both).

In both cases of thematic divergence, it is assumed that there is a CLCS that is derived from a source-language RLCS that is isomorphic to the corresponding target-language RLCS (*i.e.*, the variables in the 2 RLCS's map to the same positions, though they may be labeled differently). Furthermore, it is assumed that thematic divergence arises only in cases where there is a logical subject.

A CLCS with logical subject w, non-subject arguments $z_1, z_2, \ldots, z_k, \ldots, z_n$, and modifiers $n_1, n_2, \ldots, n_l, \ldots, n_m$ will look like the structure shown in (4), where the dominating head $\mathcal{P}$ is a typed primitive (*e.g.*, $\mathrm{BE}_{\mathrm{Circ}}$):

(4) $[\mathcal{P}\ w, z_1, z_2, \ldots, z_k, \ldots, z_n, n_1, n_2, \ldots, n_l, \ldots, n_m]$

In order to derive the syntactic structure from the CLCS, we need a mapping or *linking rule* between the CLCS positions and the appropriate syntactic positions. Roughly, this linking rule is stated as follows:

General Linking Routine $\mathcal{G}$:

(a) Map the logical subject to the external argument position.
(b) Map the non-logical-subjects to internal argument positions.
(c) Map modifiers to adjunct positions.
(d) Map the dominating head to the phrasal head position.

$\mathcal{G}$ is used for the second half of translation (*i.e.*, mapping to the target-language structure); we also need an inverse routine that maps syntactic positions of the source-language structure to the CLCS positions:

Inverse Linking Routine $\mathcal{G}^{-1}$:

(a) Map the external argument to the logical subject position.
(b) Map the internal arguments to non-logical-subject positions.
(c) Map adjuncts to modifier positions.
(d) Map the phrasal head to the dominating head node.

In terms of the representation shown in (4), the $\mathcal{G}$ and $\mathcal{G}^{-1}$ mappings would be defined as shown in figure 2.[12,13,14] Note that $w\prime$, $z_1\prime, \ldots, z_k\prime, \ldots, z_n\prime$, and $n_1\prime, \ldots, n_l\prime, \ldots, n_m\prime$ are the source-language realizations of the corresponding CLCS tokens w, $z_1, \ldots, z_k, \ldots, z_n$, and $n_1, \ldots, n_l, \ldots, n_m$; similarly, $w\prime\prime$, $z_1\prime\prime, \ldots, z_k\prime\prime, \ldots, z_n\prime\prime$, and $n_1\prime\prime, \ldots, n_l\prime\prime, \ldots, n_m\prime\prime$ are target-language realizations of the same CLCS tokens. This assumes that there is only one external argument and zero or more internal arguments. We will now look

[11] Several researchers have worked within this framework including Goldman [1974], Schank & Abelson [1977], and many others.

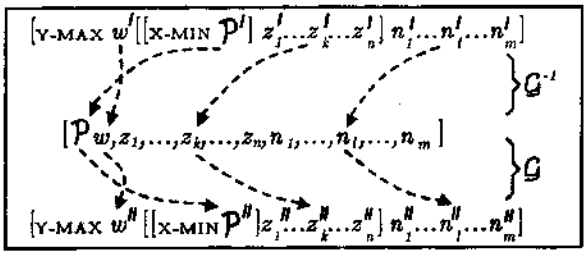

Figure 2: Mapping From Source to Target via the CLCS

at a formal description of how each type of thematic divergence is manifested. We will then see how the general linking routines described here take the syntactic mechanisms into account in order to derive the appropriate result.

4.1 Divergent Argument Positionings

In order to account for the thematic reversal that shows up in the *gustar-like* example of (1), we must have a mechanism for mapping CLCS arguments to different syntactic positions. In terms of the CLCS, we need to allow the syntactic realization of the logical subject w and the syntactic realization of a non-subject argument (say z_k) to switch places between the source and target language.

Figure 3 shows how this type of argument reversal is achieved. The :INT and :EXT markers are used in the RLCS specifications as override markers for the $\mathcal{G}$ and $\mathcal{G}^{-1}$ routines: the :INT marker is used to map the logical subject of the CLCS to an internal syntactic position (and *vice versa*). Thus, steps (a) and (b) of $\mathcal{G}$ and $\mathcal{G}^{-1}$ are activated differently if the RLCS associated with the phrasal head contains either of the :INT or :EXT override mechanisms. Note that the CLCS is the same for

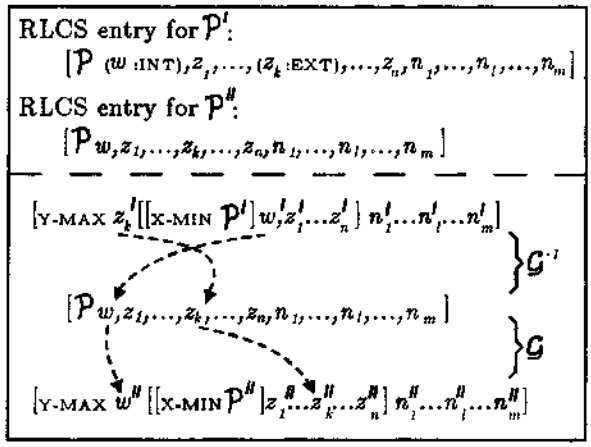

Figure 3: Mapping From Source to Target for Divergent Arguments

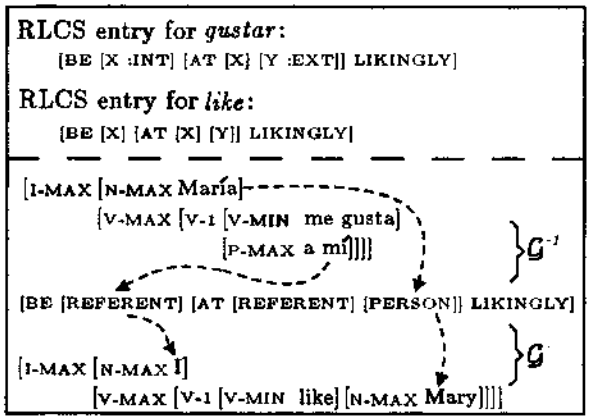

Figure 4: Translation of *María me gusta a mí*

both the source and target language; only the RLCS's in the lexical entries need to include language-specific information in order to account for thematic divergences. Now using the $\mathcal{G}$ and $\mathcal{G}^{-1}$ routines and the overriding :INT and :EXT mechanisms, we can show how to account for the thematic divergence of example (1).

Figure 4 shows the mapping from Spanish to English for example (1).[15,16] Because the Spanish RLCS includes the :INT and :EXT markers, the $\mathcal{G}^{-1}$ routine activates steps (a) and (b) differently: the external argument *María* is mapped to a non-logical-subject position [Thing PERSON], and the internal argument *mí* is mapped to the logical subject position [Thing REFERENT]. By

[12] The convention adopted in this paper is to use u' for the source-language realization, and u'' for the target-language realization for a CLCS argument u.

[13] Adjunction has been placed to the right at the maximal level. However, this is not the general case. A parameter setting determines the side and level at which a particular adjunct will occur (as discussed in [Dorr, 1990]). The configuration shown corresponds to the spec-initial/head-initial case. The other three possible configurations are: $[\text{Y-MAX}\ w'\ [\text{X-1}\ z_1'\ z_2'\ \ldots\ z_n'\ [\text{X-MIN}\ \mathcal{P}']]\ n_1', \ldots, n_m']$, $[\text{Y-MAX}\ [\text{X-1}\ [\text{X-MIN}\ \mathcal{P}']\ z_1'\ z_2'\ \ldots\ z_n']\ w'\ n_1', \ldots, n_m']$, and $[\text{Y-MAX}\ [\text{X-1}\ z_1'\ z_2'\ \ldots\ z_n'\ [\text{X-MIN}\ \mathcal{P}']]\ w'\ n_1', \ldots, n_m']$. Finally, the order of the z_i''s and n_j''s is not being addressed here; this is determined by independent principles also discussed in [Dorr, 1990]. Regardless of these syntactic variations, the $\mathcal{G}$ and $\mathcal{G}^{-1}$ routines operate uniformly because they are language-independent. For simplicity, the spec-initial/head-initial configuration will be used for the rest of this paper.

[14] In addition to realization of arguments, the dominating CLCS head ($\mathcal{P}$) must also be realized as a lexical word ($\mathcal{P}'$ in the source language and $\mathcal{P}''$ in the target language). The syntactic category of this lexical word is X, and the maximal projection is Y-MAX. In general, Y = X unless X is a Verb (in which case, Y is the Inflection category).

[15] Because of space limitations, we will illustrate the three examples (1), (2), and (3) in one direction only. However, it should be clear that the thematic divergences are solved going in the opposite direction as well since the $\mathcal{G}$ and $\mathcal{G}^{-1}$ mappings are reversible.

[16] A shorthand notation is being used for the RLCS's and the CLCS. See section 2 for a description of the actual representations used by the system.

contrast, the English RLCS does not include any special markers. Thus, the $\mathcal{G}$ routine activates steps (a) and (b) normally: the logical subject [$_{\text{Thing}}$ REFERENT] is mapped to the external argument *I*, and the non-logical-subject [$_{\text{Thing}}$ PERSON] is mapped to the internal position *Mary*.

Now we have seen how argument positioning divergences are solved during the translation process.[17] In the next section, we will look at how we account for the second part of thematic divergences: different predicate positionings.

4.2 Divergent Predicate Positionings

In the last section, we concentrated primarily on thematic interchange of *arguments*. In this section, we will concentrate on thematic interchange of *predicates*. In so doing, we will have accounted for the entire space of thematic divergences.

There are two ways to be in a predicate-argument relationship: the first is by complementation, and the second is by adjunction. That is, syntactic phrases include base-generated complements and base-generated adjuncts, both of which participate in a predicate-argument structure (where the predicate is the head that subcategorizes for the base-generated complement or adjunct).[18]

In order to show how predicate divergences are solved, we must enumerate all possible source-language/target-language predicate positionings with respect to arguments $z_1, z_2, \ldots, z_k, \ldots, z_n$ and modifiers $n_1, n_2, \ldots, n_l, \ldots, n_m$. In terms of the syntactic structure, we must examine all the possible positionings for syntactic head $\mathcal{P}$ with respect to its complements $z_1', z_2', \ldots, z_k', \ldots, z_n'$ and adjuncts $n_1', n_2', \ldots, n_l', \ldots, n_m'$.

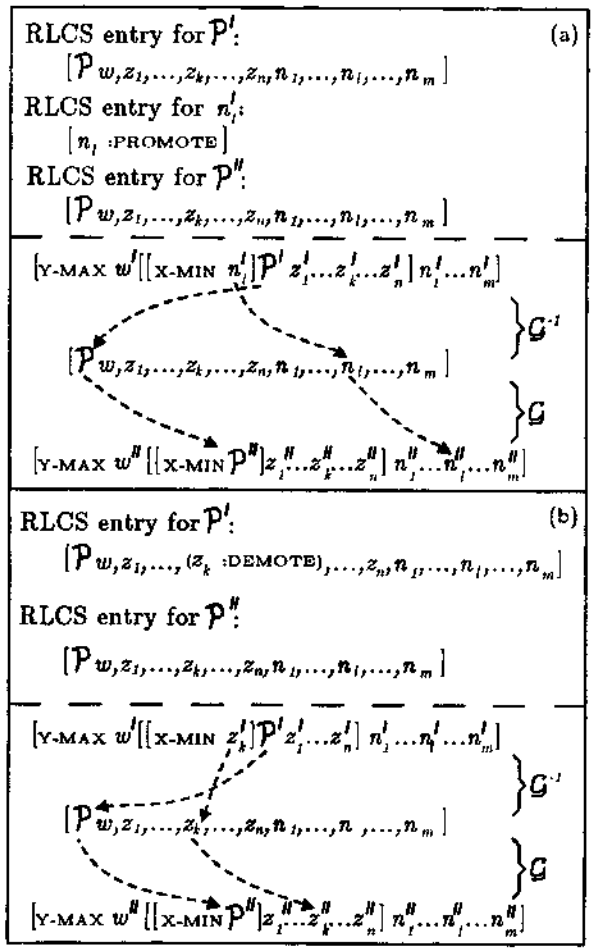

Figure 5: Mapping From Source to Target for Divergent Predicates

There are a large number of possible positionings that exhibit predicate divergences, but only two of them arise in natural language.[19] It turns out that the *soler-usually* example of (2) and the *gern-like* example of (3) are representative of the space of possibilities of predicate divergences. The source-language/target-language predicate positionings for these two cases are represented as shown in figure 5. Part (a) of this figure accounts for the translation of *usually* to *soler* (or *vice versa*), and part (b) accounts for the translation of *like* to *gern* (or *vice versa*).

The $\mathcal{G}$ and $\mathcal{G}^{-1}$ routines do not take into account the predicate divergences that were just presented. As in the case of argument divergences, predicate divergences require override markers. The :PROMOTE marker is used to map a modifier of the CLCS to a syntactic head position (and *vice versa*). The :DEMOTE marker is used to map a non-subject argument of the CLCS to a syntactic head position (and *vice versa*). Thus, steps (c), and

[17] It should be noted that the solution presented here (as well as that of the next section) does not appeal to an already-coded set of conceptual "frames." Rather, the syntactic structures are derived *procedurally* on the basis of two pieces of information: lexical entries (*i.e.*, the RLCS's) and the result of composing the RLCS's into a single unit (*i.e.*, the CLCS). It would not be possible to map *declaratively*, *i.e.*, from a set of static source-language frames to a set of static target-language frames. This is because the $\mathcal{G}$ and $\mathcal{G}^{-1}$ routines are intended to operate recursively: an argument that occurs in a divergent phrasal construction might itself be a divergent phrasal construction. For example, in the sentence *le suele gustar leer a Juan* ('John usually likes to read'), there is a simultaneous occurrence of two types of divergences: the verb *soler* exhibits a predicate positioning divergence with respect to its complement *gustar leer a Juan*, which itself exhibits an argument positioning divergence. The procedural mappings described here are crucial for handling such cases.

[18] We have left out the possibility of a base-generated specifier as a participant in the predicate-argument relationship. Of course, the specifier *is* an argument to the predicate, but it turns out that the syntactic specifier, which corresponds to the logical subject in the LCS, has a special status, and does not participate in predicate divergences in the same way as syntactic complements and adjuncts. This will be illustrated shortly.

[19] There is not enough space to elaborate on this claim here. See [Dorr, 1990] for a detailed discussion of what the possible positionings are, and which ones make sense in the context of linguistic structure.

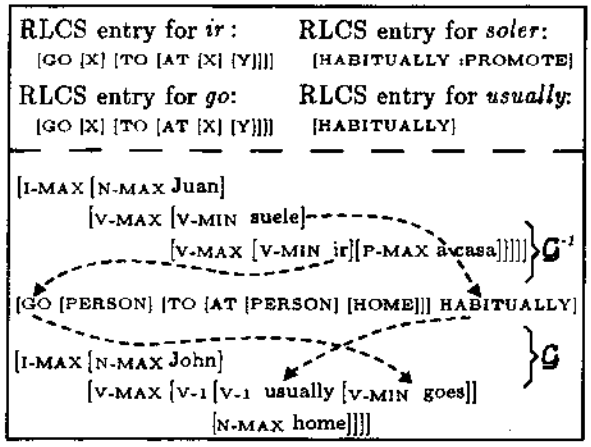

Figure 6: Translation of *Juan suele ir a casa*

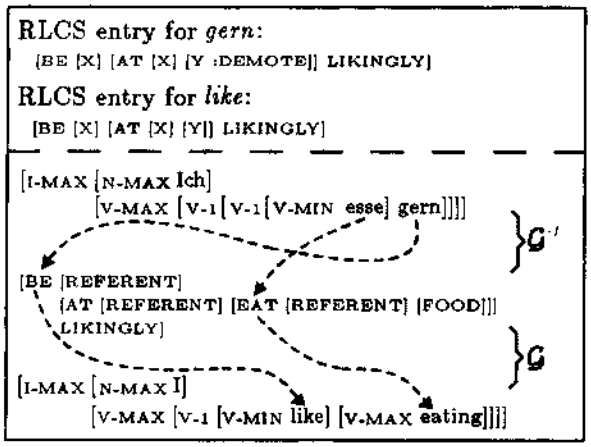

Figure 7: Translation of *Ich habe Marie gern*

(d) of the $\mathcal{G}$ and $\mathcal{G}^{-1}$ routines are activated differently if the RLCS associated with the phrasal head contains the :PROMOTE override marker, and steps (b) and (d) of these routines are activated differently if a phrasal adjunct contains the :DEMOTE override marker.

Now using the $\mathcal{G}$ and $\mathcal{G}^{-1}$ routines and the overriding :PROMOTE and :DEMOTE mechanisms, we can show how to account for the thematic divergences of examples (2) and (3) (see figures 6 and 7, respectively).

In figure 6, the Spanish RLCS for *soler* includes the :PROMOTE marker. Thus, steps (c) and (d) of $\mathcal{G}^{-1}$ are overridden: the internal argument *ir a casa* is promoted into the dominating head position [Event GO_{Loc}]; and the phrasal head *suele* is mapped into a modifier position [Manner HABITUALLY]. By contrast, the English RLCS does not include any special markers. Thus, the $\mathcal{G}$ routine activates steps (c) and (d) normally: the dominating head [Event GO_{Loc}] is mapped into the phrasal head *goes*; and the modifier [Manner HABITUALLY] is mapped into an adjunct position *usually*.

In figure 7, the German RLCS for *gern* includes the :DEMOTE marker (associated with the variable Y). Thus, steps (b) and (d) of $\mathcal{G}^{-1}$ are overridden: the phrasal head *esse* is demoted into a non-logical-subject position [Event EAT]; and the adjunct *gern* is mapped into the dominating head position [State BE_{Circ}]. By contrast, the English RLCS does not include any special markers. Thus, the $\mathcal{G}$ routine activates steps (b) and (d) normally: the dominating head [State BE_{Circ}] is mapped into the phrasal head *like*; and the non-logical-subject [Event EAT] is mapped into the internal position *eating*.

5 SUMMARY

This paper has presented a solution to the problem of thematic divergences in machine translation. The solution has been implemented in UNITRAN, a bidirectional system currently operating on Spanish, English, and German, running in Commonlisp on a Symbolics 3600 series machine. We have seen that the procedures involved are general enough to operate uniformly across different languages and divergence types. Furthermore, the entire space of thematic divergence possibilities is covered in this approach without recourse to language-specific routines or transfer rules. In addition to thematic divergences, the system handles the other divergence types shown in figure 1, and it is expected that additional divergence types will be handled by means of equally principled methods.

6 REFERENCES

[Brown, 1974] Gretchen Brown, "Some Problems in German to English Machine Translation," MAC Technical Report 142, Massachusetts Institute of Technology, Cambridge, MA, 1974.

[Chomsky, 1986] Noam A. Chomsky, *Knowledge of Language: Its Nature, Origin and Use*, MIT Press, Cambridge, MA, 1986.

[Dorr, 1987] Bonnie J. Dorr, "UNITRAN: A Principle-Based Approach to Machine Translation," AI Technical Report 1000, Master of Science thesis, Department Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[Dorr, 1990] Bonnie J. Dorr, "Lexical Conceptual Structure and Machine Translation," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1990.

[Goldman, 1974] Neil M. Goldman, "Computer Generation of Natural Language from a Deep Conceptual Base," Ph.D thesis, Computer Science Department, Stanford University, Stanford, CA, 1974.

[Jackendoff, 1983] Ray S. Jackendoff, *Semantics and Cognition*, MIT Press, Cambridge, MA, 1983.

[Lytinen & Schank, 1982] Steven Lytinen and Roger Schank, "Representation and Translation," Technical Report 234, Department of Computer Science, Yale University, New Haven, CT, 1982.

[McCord, 1989] Michael C. McCord, "Design of LMT: A Prolog-Based Machine Translation System," *Computational Linguistics*, 15:1, 33–52, 1989.

[Schank & Abelson, 1977] Roger C. Schank and Robert Abelson, *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1977.